Identifying Claims in Social Science Literature

Shameem Ahmed, Catherine Blake, Kate Williams, Noah Lenstra, and Qiyuan Liu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, USA
{ahmed9, clblake, katewill, nlenstr2, qliu14}@illinois.edu

Abstract

The Claim Framework was developed to capture how scientists communicate findings from an empirical study. Although the framework has been evaluated in biomedical literature, the framework has yet to be examined with respect to social science literature. Our goal is to fill this gap and explore the degree to which the Claim Framework can capture claims made in two social science research areas: Community Informatics and Information and Communication Technologies for Development. This poster presents preliminary results on the number and location of claims in full-text social science articles compared to claims in biomedical articles.

Keywords: claim framework, natural language processing, information extraction, community informatics, information and communication technologies for development

Introduction

Social Work Abstract Plus[1] and Sociological Abstracts[2] are two well-known databases that comprise hundreds of thousands of abstracts from thousands of social science journals. Natural Language Processing (NLP) methods have been explored to identify important concepts contained in technical papers (Paice & Jones, 1993), cause-effect relationships from news group and SIGIR datasets (Mengle & Goharian, 2010) and from newspapers (Khoo, Kornfilt, Oddy, & Hyon Myaeng, 1998) and hypotheses from MEDLINE articles (Srinivasan, 2004). However, little work has explored how well such methods will generalize to a large volume of social science abstracts. Moreover, since abstracts fails to accurately reflect the content of research articles 43 percent of the time (Pitkin, Branagan, & Burmeister, 1999), NLP methods become even more important when we start to consider full-text collections such as ERIC[3] and JSTOR[4].

In 2010, Blake proposed the Claim Framework (Blake, 2010), as a domain-independent representation of how scientists communicate their findings in empirical studies. The framework defines claim as new finding from the articles that brings about an effect or a result. For instance, "*Indeed, glycine prevented Wy-14643-stimulated superoxide production by Kupffer cells*" is a claim in biomedical literature. On the other hand, "*Contrary to much rhetoric, even very poor people chose to have a phone*" is a claim collected from social science literature. Although the Claim Framework was developed for the life sciences, such as bioinformatics and clinical informatics, it is not clear how well the framework will generalize to findings reported in the social sciences literature.

Our goal is to explore the extent to which claims made by authors in the social sciences conform to the Claim Framework (Blake, 2010). This poster describes the first step towards that goal by identifying claims in eight full-text articles in two social science research domains: Community Informatics (CI) (Gurstein 2000; Keeble & Loader, 2001; Williams & Durrance, 2009) and Information and Communication Technologies for Development (ICT4D) (Unwin, 2009).

[1] www.ovid.com/site/catalog/DataBase/150.jsp
[2] http://www.csa.com/factsheets/socioabs-set-c.php
[3] www.eric.ed.gov
[4] www.jstor.org

Ahmed, S., Blake, C., Williams, K., Lenstra, N., & Liu, Q. (2013). Identifying claims in social science literature. *iConference 2013 Proceedings* (pp. 942-946). doi:10.9776/13485

Method

Eight full-text peer-reviewed research articles from two journals of CI and ICT4D (Journal of Community Informatics and Information Technology for Development Journal) were collected. These articles were selected at random from a dataset collected in our earlier study (Williams, Ahmed, Lenstra, & Liu, 2012). Scripts were written to segment the documents into sentences and segmentation errors were corrected. Although different articles contained different number of sections, all eight articles comprised at least five sections, namely, Abstract, Introduction, Research Method, Results, and Conclusion. To standardize, additional sections present in each paper were moved into one of the five categories. Each sentence was annotated either as a claim or as a non-claim.

Results and Discussion

A total of 2,433 sentences remained after sentence errors such as missing spaces after a period, unknown characters, typographical errors, and ambiguous references were corrected. Table 1 summarizes the number of claims made in each article.

Table 1
Claims in CI and ICT4D articles

		Abstract	Introduction	Method	Results	Conclusion	Total
CI1	Number of sentences	3	57	17	336	51	464
	Number of claims	0	0	0	136	30	166
CI2	Number of sentences	4	82	7	170	21	284
	Number of claims	3	3	0	110	11	127
CI3	Number of sentences	7	36	22	137	28	230
	Number of claims	7	2	0	60	16	85
CI4	Number of sentences	5	13	99	60	170	347
	Number of claims	1	0	14	41	93	149
ICT4D1	Number of sentences	6	62	37	153	29	287
	Number of claims	5	5	0	81	20	111
ICT4D2	Number of sentences	5	25	19	206	24	279
	Number of claims	4	7	2	38	8	59
ICT4D3	Number of sentences	6	71	46	31	64	218
	Number of claims	3	5	1	19	45	73
ICT4D4	Number of sentences	10	132	24	121	37	324
	Number of claims	5	1	0	48	30	84

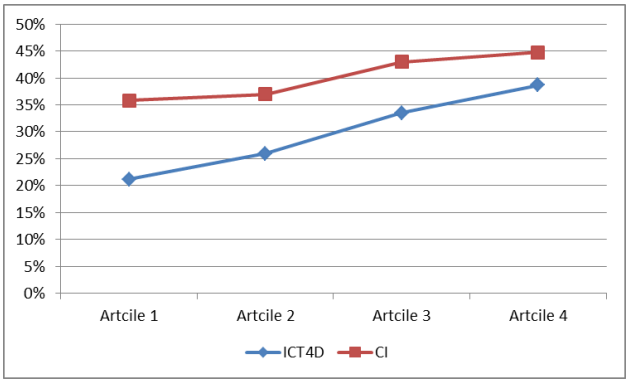

Figure 1. Percentage of Claims for CI and ICT4D articles

Table 2
Comparative Claims for CI and ICT4D articles

	CI	ICT4D
Average % of claims per article	40.1%	29.8%
Minimum % of claims per article	35.8%	21.1%
Maximum % of claims per article	44.7%	38.7%

Figure 1 shows the percentage of claims and Table 2 summarizes the comparative claim statistics for CI and ICT4D articles. The percentage of claims is calculated by dividing the number of sentences that report a claim by the total sentences in an article (excluding references). For example, if there are 100 sentences in an article and among those sentences 34 are considered as claim sentences then percentage of claim for that article is 34%.

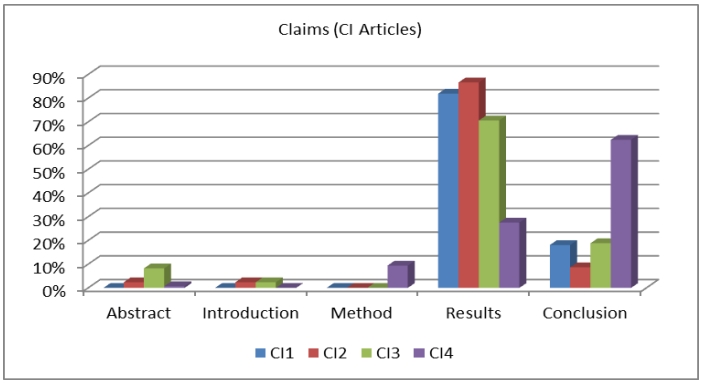

Figure 2. Claims for CI articles

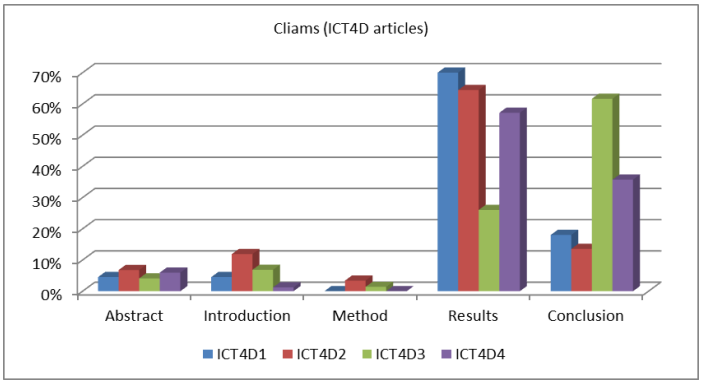

Figure 3. Claims for ICT4D articles

Figure 2 and 3 show that the distribution of claims between sections is similar for both CI and ICT4D articles. Here, we see that, authors frequently report claims in the Results and Conclusion sections. However, authors in CI report more results in the Results section than in ICT4D (66.7% vs. 55.1%) and authors of ICT4D articles report more results in the Conclusion section than CI (32.3% vs. 27.0%). The contribution of other three sections, both for ICT4D and CI, are negligible with respect to

claims. These results suggest that Text Mining systems should have more focus on Results and Conclusion section for ICT4D and CI articles, which is quite different than biomedical research domain.

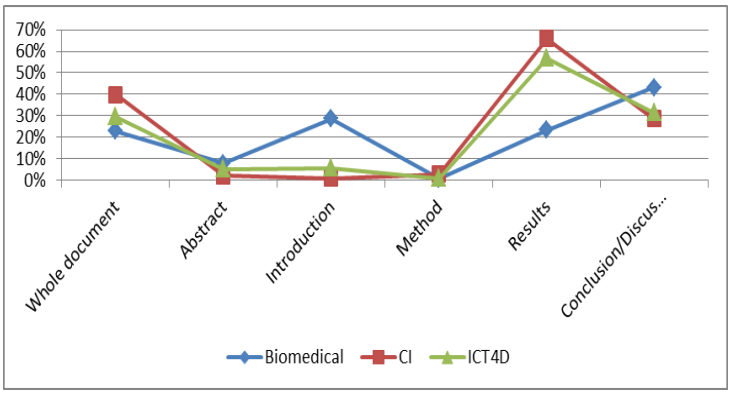

Figure 4. Claims for Biomedical vs. CI vs. ICT4D articles

As the Claim Framework has already been tested in biomedical articles, we can provide a side-by-side comparison for claims made in biomedical, CI, and ICT4D articles. Figure 4 shows that the proportion of claims made per article is higher in CI and ICT4D articles than in biomedical articles (39.8% vs. 29.5% vs. 22.8%). Authors in biomedical research area tend to have a greater proportion of claim sentences in the abstract section than in social science literature (7.8% vs. 2.1% vs. 5.2%). In addition, authors of biomedical articles have a greater proportion of claim sentences (28.6%) in the Introduction section. In contrast, CI has only 0.9% and ICT4D has 5.5% claims in the Introduction. It seems that while CI and ICT4D authors mainly discuss their motivation, research question, and some background information in the Introduction section, biomedical authors report one-third of their claims in this section.

Another difference was found with respect to the document structure. Only 2 out of the 29 biomedical articles included a conclusion section, whereas all 8 articles in the CI and ICT4D articles included a conclusion section. Sentences in the discussion section of a biomedical article provide the reader with information about context and the implications of the study results, which are important, but differ from the focus of the Claim Framework which includes factual statements about the study findings. In the CI and ICT4D articles, a greater proportion of the sentences in the Result section report claims. As expected, none of the research communities focus on claims in the Methods section (0.5% vs. 2.7% vs. 0.9% for CI, ICT4D, and Biomedical respectively).

Conclusion

Several attempts to characterize scientific literature have been made in the life sciences, but little work has been done to explore how well those methods might apply to the social sciences. Our goal is to explore the extent to which Blake's Claim Framework (Blake, 2010) might apply to social science articles.

Although we have yet to conduct a more detailed analysis of how claims identified in these articles intersect with the five Claim Framework types (explicit, implicit, comparison, correlation, and observation), these preliminary results suggest that there may be differences between social science literature and the biomedical literature with respect to where an author is likely to report study results (claims). Findings reported in this paper have important implications for both information retrieval and natural language processing systems. We plan to increase the number of articles in the analysis as future work and to apply automated methods to identify explicit claims and comparisons described in (Blake, 2010) and (Hoon Park & Blake, 2012) respectively.

References

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2), 173-189.

Gurstein, M. (2000). Community informatics enabling communities with information and communications technologies. Hershey, Pa. Idea Group Pub.

Hoon Park, D, & Blake, C. (2012). Identifying comparative sentences in full-text scientific articles. *Association of Computational Linguistics, Workshop on Detecting Structure in Scholarly Discourse*, Jeju South Korea, 2012.

Keeble, L., & Loader, B. (2001). Community informatics : shaping computer-mediated social relations. Routledge, New York.

Khoo, C., Kornfilt, J., Oddy, R. N., & Hyon Myaeng, S. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, *13(4)*, 177-186.

Mengle, S. S. R. & Goharian, N. (2010). Detecting relationships among categories using text classification. *Journal of the American Society for Information Science and Technology*, *61*, 1046–1061.

Paice, C. D., & Jones, P. A. (1993). The identification of important concepts in highly structured technical papers. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 69-78.

Pitkin, R. M., Branagan, M. A., & Burmeister, L. F. (1999). Accuracy of Data in Abstracts of Published Research Articles. *The Journal of American Medical Association, 281(12)*, 1110-1111.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55*, 396–413.

Unwin, T. (2009). ICT4D: Information and Communication Technology for Development. Cambridge University Press.

Williams, K., Ahmed, S., Lenstra, N., & Liu, Q. (2012). What is Community Informatics? A Global and Empirical Answer. *iConference*, Toronto, Canada, 2012.

Williams, K., & Durrance, J. C. (2009). Community Informatics. *Encyclopedia of Library and Information Science*, M. Bates & M. Miles Maack (Ed).